DOCUMENT RESUME

ED 334 212 TM 016 633

AUTHOR Yoes, Michael E.; Ho, Kevin T.
TITLE The Degree of Person Misfit on a Nationally Standardized Achievement Test.
PUB DATE Apr 91
NOTE 34p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, IL, April 4-6, 1991).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Achievement Tests; Comparative Testing; Elementary Secondary Education; Estimation (Mathematics); *Goodness of Fit; National Programs; Reading Comprehension; Reading Tests; Science Tests; Simulation; Spelling; *Standardized Tests; Testing Programs; *Test Interpretation
IDENTIFIERS Ability Estimates; *Person Fit Measures; Stanford Achievement Tests

ABSTRACT

How two students with the same number-correct score, calculated by dramatically different patterns of responses, can achieve the same ability estimate is demonstrated. The present study investigated the incidence of person misfit to the model in a nationally standardized test of achievement devised using the Rasch item response theory model. It was hypothesized that if guessing occurs frequently and the Rasch model is not robust enough to handle it, then the percentage of examinees with aberrant response patterns (i.e., persons misfitting the model) should be high. Using student data for grades 3 through 12 from the science, reading comprehension, and spelling subtests of the spring 1988 administration of the Stanford Achievement Test (STAT) (Eighth Edition), response patterns were analyzed with three indicators of person fit (unweighted Rasch mean square fit, INFIT, and standardized likelihood). Data for 45,442 examinees on the science subtest, 47,040 examinees on the reading comprehension subtest, and 46,332 examinees on the spelling subtest were analyzed for each of the STAT and simulated data sets. The percentages of students in the national standardization sample identified by each indicator and by all three as potentially misfitting the Rasch model are presented along with a baseline analysis of model-fitting simulation data. Results of this study indicate that the percentage of students identified as misfitting the model is small. While the utility and choice of person fit indicators may be arguable, the data show the usefulness of the Rasch model for educational applications. Seven data tables and five graphs are included. (RLC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED334212

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

MICHAEL E. YOES

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

The Degree of Person Misfit on a Nationally Standardized Achievement Test

Michael E. Yoes

The Psychological Corporation

Kevin T. Ho

Jostens Learning Corporation

Paper presented at the Annual Meeting
of the National Council on Measurement in Education
Chicago, April 1991

TM016633

BEST COPY AVAILABLE

Abstract

One of the major advantages of the Rasch IRT model has also been expressed as one of its criticisms, that is, of course, the one-to-one correspondence between number-correct score and the IRT estimate of ability/achievement. A demonstration is offered to show how two students with the same number-correct score, arrived at by dramatically different patterns of responses, can achieve the same ability estimate. To the extent that this type of response behavior happens often, the application of the Rasch model would be questionable. The present study was designed to investigate the incidence of person misfit to the model in a nationally standardized test of achievement built using the Rasch IRT model. Using student data (Grades 3 - 12), from the Science, Reading Comprehension, and Spelling subtests of the Stanford Achievement Test, Eighth Edition, response patterns were analyzed using three indicators of person fit (Rasch Mean Square Fit, INFIT, and Standardized Likelihood). The percentages of students in the national standardization sample identified by each indicator (and by all three) as potentially misfitting the Rasch model are presented along with a baseline analysis of model-fitting simulation data. Results of this study appear to indicate that the percentage of students identified as misfitting the model is small. While the utility, and choice, of person fit indicators may be arguable these data do offer evidence of the usefulness of the Rasch model for educational applications.

The Degree of Person Misfit on a Nationally Standardized Achievement Test

Introduction

Item response theory (IRT) has revolutionized the field of educational measurement. Applications of IRT to item banking, equating, and test development are in widespread use (Hambleton and Swaminathan, 1985; Hambleton, 1989). The most parsimonious of the IRT models used by practitioners involved with educational assessments is the one-parameter logistic (Rasch) model that allows items to differ only in terms of their difficulty level. In many practical (e.g., educational) settings, the Rasch model is the only viable IRT model because it requires much smaller sample sizes, than other IRT models, in order to obtain stable item parameter estimates.

One of the major advantages of the Rasch model (particularly with group administered tests) has also been expressed as one of its criticisms; that is, of course, the one-to-one correspondence between number-correct score and the IRT estimate of ability/achievement level. In practical testing situations, the ability to hand score test materials and/or to verify machine-scored answer documents is a desirable feature. Critics of the Rasch model assert that the model is too simplistic and fails to adequately take into account the possibility of guessing on the part of examinees. They point out that when violations of the model occur, particularly in the form of guessing, it is possible for two students to obtain the same number-correct score by different patterns of responses

to the test items. For example, on a ten-item test, two students (student A and student B) both obtain a number-correct score of five. Student A answers the five easiest items correctly while missing the five most difficult and exemplifies a perfect fit to the model. Student B answers the five most difficult items correctly, misses the five easiest items, and exemplifies a misfit to the model. While the response patterns of students A and B represent the extremes in terms of person fit to the Rasch model, it is likely that many types of response patterns would arise from this ten-item test. If there are many students whose response patterns suggest misfit to the model, then either the robustness of the Rasch model or the characteristics of the student population needs to be examined.

Measures of person fit (Hulin, Drasgow and Parsons, 1983; Wright and Stone, 1979; Wright and Masters, 1982) that indicate whether or not individuals are responding to the test items in a manner consistent with the Rasch model can be used to examine the percentage of students not responding according to expectations. The Rasch IRT model conceptualizes person fit as a process of identifying "unexpected" responses by examinees to test items. The basis of classifying a response by an examinee to an item as "unexpected" is quantified based on expectations prescribed by the model. Typically fit indicators begin by examining the difference between an examinee's ability level and the item difficulty level ($\theta - b$), since people and items are expressed on the same scale. When the difference between ability and item difficulty is positive (when an individual's ability level exceeds the item difficulty) the item should be easy for that person. When the

difference ($\theta - b$) is negative (i.e., when the item difficulty is greater than the person ability) the item becomes more difficult for the examinee.

Surprisingly, there appears to have been little or no empirical examination of the amount, or incidence, of person misfit in large-scale implementations of the Rasch model. Instead, most investigations have focused on the evaluation of the various indicators of person fit themselves rather than their application. Additionally, many of these studies have relied almost exclusively on simulation data in their evaluations.

The present investigation seeks to examine the incidence rate of person misfit, as evidenced by three currently available techniques, in data obtained from the standardization of a nationally recognized achievement test that was developed using the Rasch model. If guessing occurs frequently and the Rasch model is not robust enough to handle it, then the percentage of examinees with aberrant response patterns (i.e., persons misfitting the model) should be high.

Instruments Used

The Stanford Achievement Test Series, Eighth Edition is designed to reflect what is being taught in schools throughout the country. The test series is the result of a thorough analysis of curriculum materials as well as input from curriculum specialists. Testing for the standardization of the Stanford series was conducted in the spring and

fall of 1988 with over 200,000 nationally representative students in grades Kindergarten through 12 participating. The Stanford Achievement Test Series, Eighth Edition was developed using a Rasch IRT model, and is well respected for its validity and reliability.

Three subtests from the Stanford Achievement Test Series were selected for use in the present study: Science, Reading Comprehension, and Spelling. It should be noted that the Stanford Achievement Test Series consists of test batteries whose 13 levels span grades Kindergarten through twelve. Not all subtests are included at all levels of the Stanford and also the length of the subtests can vary across levels. The three selected subtests (Science, Reading Comprehension, and Spelling) were specially chosen for this investigation because they were included in the Stanford battery at all of the grades (3-12) and because the length of these subtests are relatively constant across levels. Table 1 presents the Stanford test levels, by grade, and subtest lengths for the three subtests used in this investigation.

Description of Subtests

Science. The Science subtest of the Stanford Test Series is built upon the philosophy that the biological, physical, and earth/space sciences form the core of the science curriculum, with the science process skills forming the basis for understanding and evaluating science content. The premise endorsed by the subtest is one of "scientific literacy", with the major emphasis being on the concepts of science rather than on

isolated facts. All science items are in a multiple-choice, 4-option, format.

Reading Comprehension. The Reading Comprehension subtest of the Stanford in accordance with research indicating that comprehension may be affected by the type of text being read, assesses reading comprehension with three types of reading material (textual, recreational, and functional). Textual content is similar to that found in nonfiction, general information materials. Recreational content includes material read for enjoyment, including fiction, humor, literary criticism, and poetry. Functional content is the type of material encountered in everyday life, e.g., directions, forms, advertising, and labels. Within each type of reading material, test items are constructed to tap literal, inferential, and critical levels of comprehension. Each reading passage has associated with it several 4-option (multiple-choice) test items.

Spelling. The Spelling subtest of Stanford assesses objectives based upon the phonetic and structural principles taught in school. Phonetic and structural principles are tested with four-option (multiple-choice) items that include three correctly spelled words and one misspelled word. The misspellings in this subtest reflect the most plausible and commonly made errors in the spelling of that word.

Methods

Subjects

Data for this investigation were collected as part of the spring 1988 national standardization of the Stanford Achievement Test Series. All analyses were conducted on student response data resulting from testing students in grades 3 through 12 with the grade appropriate level of the Stanford test battery. As an additional edit criteria examinees that earned zero number-correct scores or that omitted 10% or more of the items in any of the subtests were deleted from all analyses for that subtest.

Computation of Person Fit

Typical measures of person fit with the Rasch model focus on standardized residuals, each residual is computed based on the difference between the examinee's obtained response (0 = incorrect, 1 = correct) and the model predicted probability of a correct response. These summed (standardized) residuals are then divided by an estimated standard error term.

Person fit in the present investigation was analyzed using three indicators (described individually below): (1) an unweighted standardized mean square (Mean Square Fit -- Wright and Stone, 1979), (2) a standardized weighted mean square procedure (INFIT) described by Wright and Masters (1982), and (3) a standardized measure based on the natural logarithm of the maximum of the likelihood function (Drasgow, Levine, and Williams, 1982).

Data from each of the three Stanford subtests (science, reading comprehension, and spelling) were analyzed for person fit using each of the three indices (significance level of $\alpha = 0.05$). A computer program written for this purpose computed the three fit indices. Rasch item difficulties based on the standardization data and maximum-likelihood estimates of ability for each examinee were used in combination with student response data in order to determine fit of examinees to the model. Results of these analyses were summarized to report the percentages of the national standardization sample (by grade) that were identified by each index as potentially misfitting the Rasch IRT model. The percentage of examinees identified by all three fit measures as potentially misfitting the model was also obtained.

Person Fit Measures Used

Mean Square Fit. The unweighted Mean Square Fit index (hereafter referred to as MS Fit) is described in Wright and Stone (1979) and is the simplest of the Rasch person fit indicators. This measure is an average of the squared standardized residuals for that person over the n items to which that examinee responded. This indicator reportedly is distributed as an F statistic with $(n - 1)$ and ∞ degrees of freedom. It is the typical measure of fit (person and item) reported in the 'older' Rasch software (e.g., BICAL). Critical values ($\alpha = 0.05$) of the F distribution for degrees of freedom $(n - 1)$ and ∞ were used for each subtest by grade (level of test) combination. Students whose MS Fit

value exceeded the critical value were flagged as potentially misfitting the Rasch model.

INFIT. One of the problems that was noted early on when using the unweighted MS Fit index is that it can be overly sensitive to unexpected responses made by an examinee to items that were far too easy or too difficult. When using the unweighted MS Fit index it is possible to identify individuals as misfitting the Rasch model on the basis of only two or three unexpected responses. To compensate for this occurrence Wright and Masters (1982) outlined a weighted mean square fit statistic such that responses made by an examinee to items that are further away from that examinee's ability level will have less weight than responses in the vicinity of the examinee's ability. Wright and Masters (1982) provide a transformation of the weighted mean square fit measure into an index referred to as INFIT. When the data fit the model the INFIT measure has a mean near zero and a standard deviation near one. Newer versions of Rasch software utilize the INFIT measure as an indicator of model fit. Students in the present investigation whose INFIT measure either exceeded +2.0 or was less than -2.0 were flagged as potentially misfitting the Rasch model.

Standardized Likelihood. As was noted previously, individuals receiving the same number-correct score will receive the same maximum-likelihood ability estimate when the Rasch IRT model is used, regardless of their pattern of responses. There may, however, be dramatic differences in the height of the likelihood function at that ability value.

Consider the two examinees mentioned in the introductory section of this paper, each received a number-correct score of 5 but Student A attained that score by answering the 5 easiest items correctly and Student B attained a score of 5 by answering the five most difficult items correctly. Clearly Student A perfectly fits the model and Student B is not responding in accordance with the expectations of the model. Figure 1 compares the likelihood functions obtained by Student A and Student B given their respective patterns of responses.

Insert Figure 1 about here

Insert Figure 2 about here

Both response patterns in Figure 1 would be assigned the same maximum-likelihood estimate of ability when the Rasch model is used. The likelihood function for Student B appears flat compared to the likelihood function for Student A but, in fact, its appearance simply reflects a tremendous difference in the magnitude of the two functions. Figure 2 demonstrates this difference by comparing the two likelihood functions after scaling each of them by dividing each ordinate by the maximum of that function. As can be seen in figure 2, the shape and peak of the functions are identical when differences in scale are taken into account. It would appear, therefore that the height of the likelihood function would have utility as a measure of evaluating the fit of an examinee to the model.

Drasgow, Levine, and Williams (1982) provided a measure of person fit based on the natural logarithm of the likelihood function at its maximum. This measure is standardized to express it in relation to an expected likelihood value (given the model) and an approximate standard deviation. The resulting index, referred to in this paper as the standardized likelihood [l_z], is reported to follow a normal (0,1) distribution. Students in this investigation whose l_z measure exceeded +2.0 or was less than -2.0 were flagged as potentially misfitting the Rasch model.

Analysis of Simulation Data

Model generated simulation data were also analyzed in order to provide a baseline from which to interpret the performance of each of the person fit measures. Simulation data for each of the grade level combinations were generated using the standardization Rasch item difficulty values and the standardization examinee maximum-likelihood estimates of ability. These Rasch-model simulated data files were then analyzed using the same procedures as the actual Stanford standardization data.

Results/Conclusions

In order to facilitate the explanation of the results all summaries are in reference to misfitting examinees rather than examinees that fit the model. The reader should keep in mind, however, that for each of the tables presented there is an analogous table that could report the

percentages of examinees that did respond in accordance with the model, as measured by the three fit indices used in this investigation.

Table 2 reports the sample sizes and percentage of the spring 1988 Stanford standardization sample, by grade, identified by each fit indicator as potentially misfitting the Rasch model based on data from the Science subtest. As can be seen in the table the percentages are generally small. For the standardized likelihood [l_z] and INFIT measures, all percentages are within expectations based on the specified significance level of 0.05. Tables 3 and 4 report similar results for the Reading Comprehension and Spelling subtests of the Stanford. As can be seen in all of the tables the percentage of misfitting students was generally low, particularly for the INFIT and standardized likelihood measures. Also note that the percentage of misfitting students does not increase with age as one might expect with increasing "test wiseness" and/or increased guessing behaviors.

Insert Table 2 about here

Insert Table 3 about here

Insert Table 4 about here

Note that the percentages of students identified as potentially misfitting the model using the MS Fit measure are generally greater than the other two measures. This finding is consistent with the

aforementioned sensitivity of this measure to small numbers of "unexpected" responses. The percentage of the spring 1988 Stanford standardization sample, by grade, identified by all three indices as potentially misfitting the model was also very small. Note that the percentages of students identified as misfitting the model appears to be highest at Grade 7, at least for Reading Comprehension and Spelling (Normative data for Stanford and other nationally standardized achievement tests typically show a decrement in performance at grade 7). Figures 3 through 5 present a graphic representation of these same results.

Insert Figure 3 about here

Insert Figure 4 about here

Insert Figure 5 about here

Results of the analyses of the simulation data are reported in Tables 5 through 7 and is offered as a baseline measure to aid in the interpretation of the results involving the standardization data. As can be seen in these tables the MS Fit measure appears to be the most conservative measure (i.e., flagging larger percentages of 'simulees' as misfitting the model). The standardized likelihood and INFIT measures appear to function much better than does the MS Fit measure. Based on these findings it is recommended that caution be used in the application of the more conservative MS Fit index as a measure of person fit.

Insert Table 5 about here

Insert Table 6 about here

Insert Table 7 about here

Although the psychometric properties for measures of person fit continue to be argued, and investigated, the results from this investigation do provide further support that, when used properly, the Rasch model does work in educational settings. Based on the relatively low percentage of students in the Stanford national standardization sample, identified as potentially misfitting the Rasch model, the criticisms of the robustness of the Rasch model would appear to be unfounded.

References

Drasgow, F., Levine, M.V., and WIlliams, E. 1982. Advances in appropriateness measurement. Cited from Hulin, C.H., Drasgow, F., and Parsons, C.K. 1983. Item Response Theory: Application to Psychological Measurement. Homewood, IL: Dow Jones-Irwin.

Hambleton, R.K. 1989. Principles and selected applications of Item Response Theory. In R. Linn (Ed.) Educational Measurement, 3rd Edition. Washington, D.C.: National Council on Measurement in Education and the American Council on Education.

Hambleton, R.K. and Swaminathan, H. 1985. Item Response Theory: Principles and Applications. Boston, MA: Kluwer Academic Publishers.

Hulin, C.H., Drasgow, F., and Parsons, C.K. 1983. Item Response Theory: Application to Psychological Measurement. Homewood, IL: Dow Jones-Irwin.

Wright, B.D., and Masters, G.N. 1982. Rating Scale Analysis. Chicago: Mesa Press.

Wright, B.D., and Stone, M.H. 1979. Best Test Design. Chicago: Mesa Press.

Tables

Table 1. Levels of the Stanford Achievement Test, Eighth Edition and Subtest Lengths corresponding to Grade Level

GRADE	Stanford Level	Subtest Length		
		Science	Reading Comprehension	Spelling
3	Primary 3	50	54	36
4	Intermediate 1	50	54	40
5	Intermediate 2	50	54	50
6	Intermediate 3	50	54	50
7	Advanced 1	50	54	50
8	Advanced 2	50	54	50
9	TASK 1	50	54	40
10	TASK 2	50	54	40
11	TASK 3	50	54	40
12	TASK 3	50	54	40

Table 2. Percent of SAT8 Spring Standardization Sample Identified as Misfitting the Rasch Model

($\alpha = 0.05$)

Science Subtest

GRADE	Sample Size	Person Fit Indices			
		l_z	INFIT	MS Fit	ALL Three
3	5,158	4	4	8	2
4	5,422	4	4	5	1
5	5,852	5	4	8	2
6	5,763	4	4	5	1
7	5,443	4	4	6	2
8	4,961	2	2	2	< 1
9	2,294	3	2	3	1
10	3,959	3	3	3	1
11	3,698	3	3	4	1
12	2,892	3	3	4	< 1

Table 3. Percent of SAT8 Spring Standardization Sample Identified as Misfitting the Rasch Model

($\alpha = 0.05$)

Reading Comprehension Subtest

GRADE	Sample Size	Person Fit Indices			
		l_z	INFIT	MS Fit	ALL Three
3	5,329	3	3	6	1
4	5,702	5	4	7	2
5	6,068	4	3	4	1
6	5,837	4	4	7	2
7	5,461	7	6	10	4
8	5,020	3	3	5	1
9	2,353	6	5	8	2
10	4,183	5	4	9	3
11	3,954	3	3	5	1
12	3,133	4	3	6	1

Table 4. Percent of SAT8 Spring Standardization Sample Identified as Misfitting the Rasch Model

($\alpha = 0.05$)

Spelling Subtest

GRADE	Sample Size	Person Fit Indices			
		l_z	INFIT	MS Fit	ALL Three
3	5,497	4	4	6	1
4	5,740	4	3	8	2
5	5,705	5	5	10	3
6	5,491	3	4	7	1
7	5,336	5	5	7	2
8	4,903	4	4	8	2
9	2,381	4	4	9	2
10	4,199	4	4	10	2
11	3,937	4	4	5	1
12	3,143	4	3	7	1

Table 5. Percent of Rasch Model Simulated Data Identified as Misfitting the Rasch Model

($\alpha = 0.05$)

Science Subtest

GRADE	Sample Size	Person Fit Indices			
		l_z	INFIT	MS Fit	ALL Three
3	5,158	1	1	7	< 1
4	5,422	2	2	3	< 1
5	5,852	2	2	6	< 1
6	5,763	2	2	5	< 1
7	5,443	1	1	4	< 1
8	4,961	1	1	3	< 1
9	2,294	1	1	2	< 1
10	3,959	2	2	1	< 1
11	3,698	2	2	2	< 1
12	2,892	1	1	1	< 1

Table 6. Percent of Rasch Model Simulated Data Identified as Misfitting the Rasch Model

($\alpha = 0.05$)

Reading Comprehension Subtest

GRADE	Sample Size	Person Fit Indices			
		l_z	INFIT	MS Fit	ALL Three
3	5,329	1	1	5	< 1
4	5,702	1	2	5	< 1
5	6,068	1	2	4	< 1
6	5,837	1	2	5	< 1
7	5,461	2	2	7	< 1
8	5,020	1	1	3	< 1
9	2,353	1	1	3	< 1
10	4,183	1	2	6	< 1
11	3,954	1	2	4	1
12	3,133	1	2	4	< 1

Table 7. Percent of Rasch Model Simulated Data Identified as Misfitting the Rasch Model

($\alpha = 0.05$)

Spelling Subtest

GRADE	Sample Size	Person Fit Indices			
		l_z	INFIT	MS Fit	ALL Three
3	5,497	1	1	5	< 1
4	5,740	1	1	7	< 1
5	5,705	2	3	8	1
6	5,491	2	2	5	< 1
7	5,336	1	2	5	< 1
8	4,903	1	2	6	< 1
9	2,381	2	2	6	1
10	4,199	2	2	8	1
11	3,937	1	1	4	< 1
12	3,143	1	1	5	< 1

Figures

Figure 1. Comparison of Original Likelihood Functions.

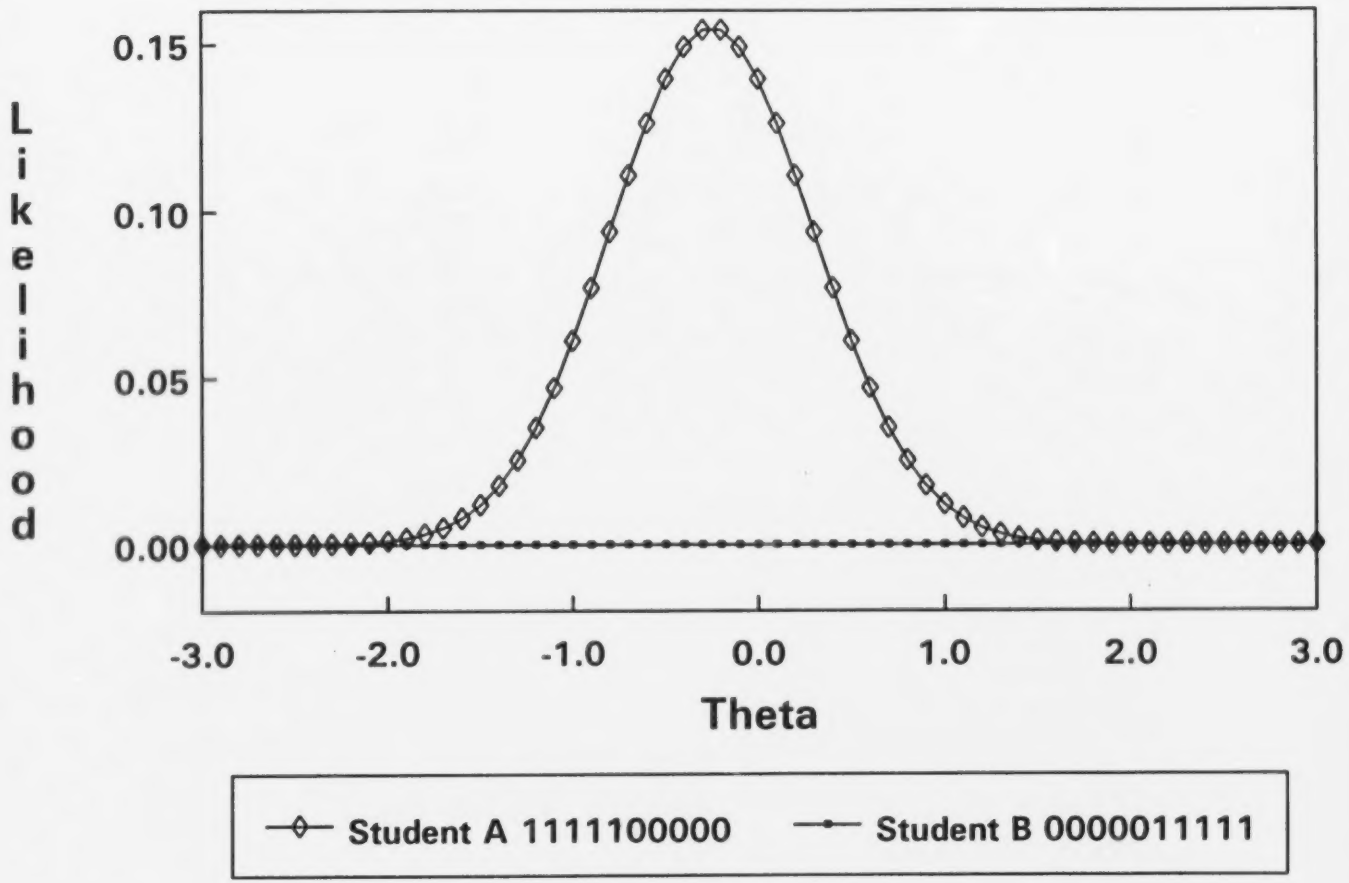


Figure 2. Comparison of Scaled Likelihood Functions.

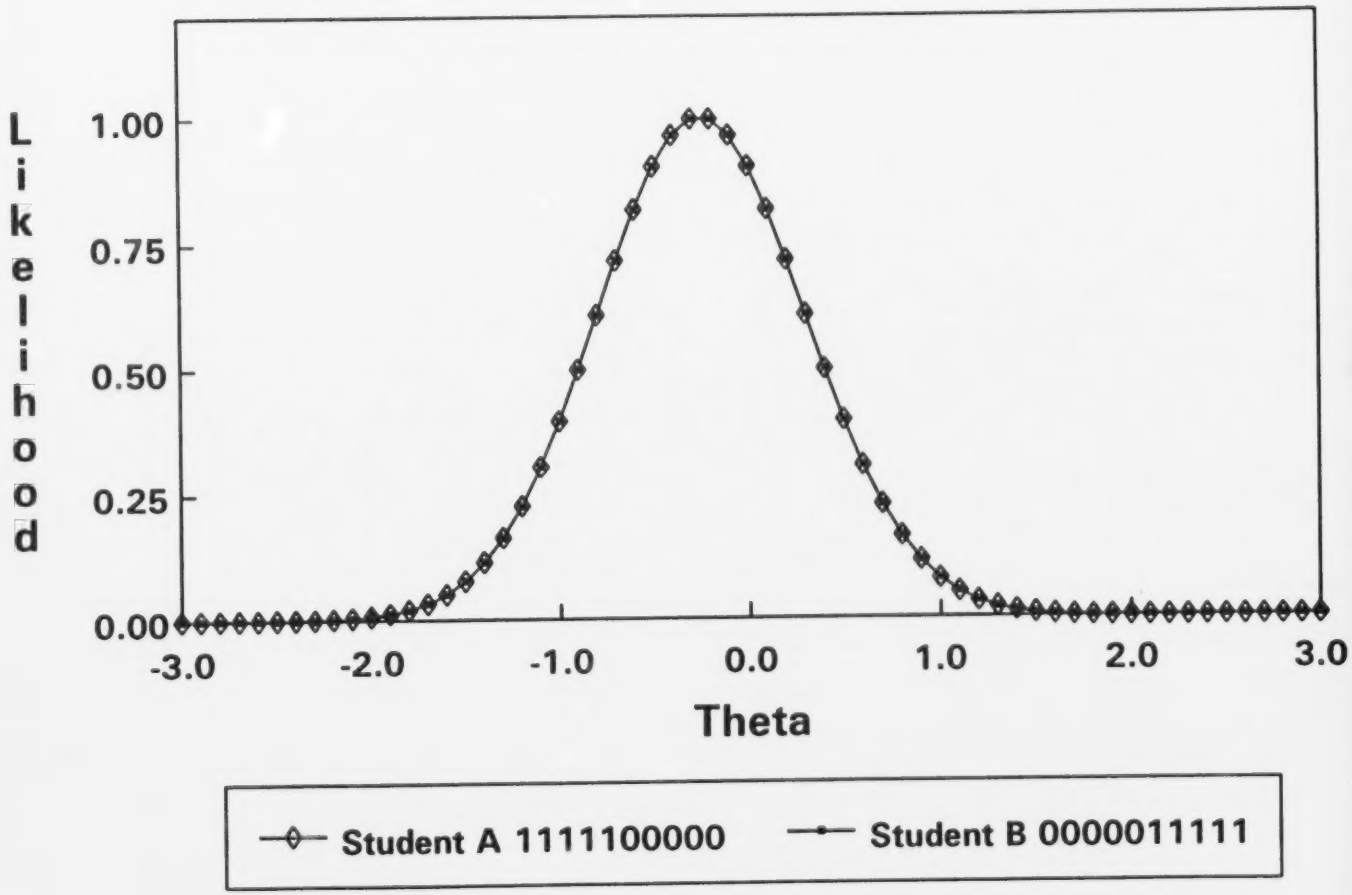


Figure 3
Percentage of Misfitting Students
Science Subtest

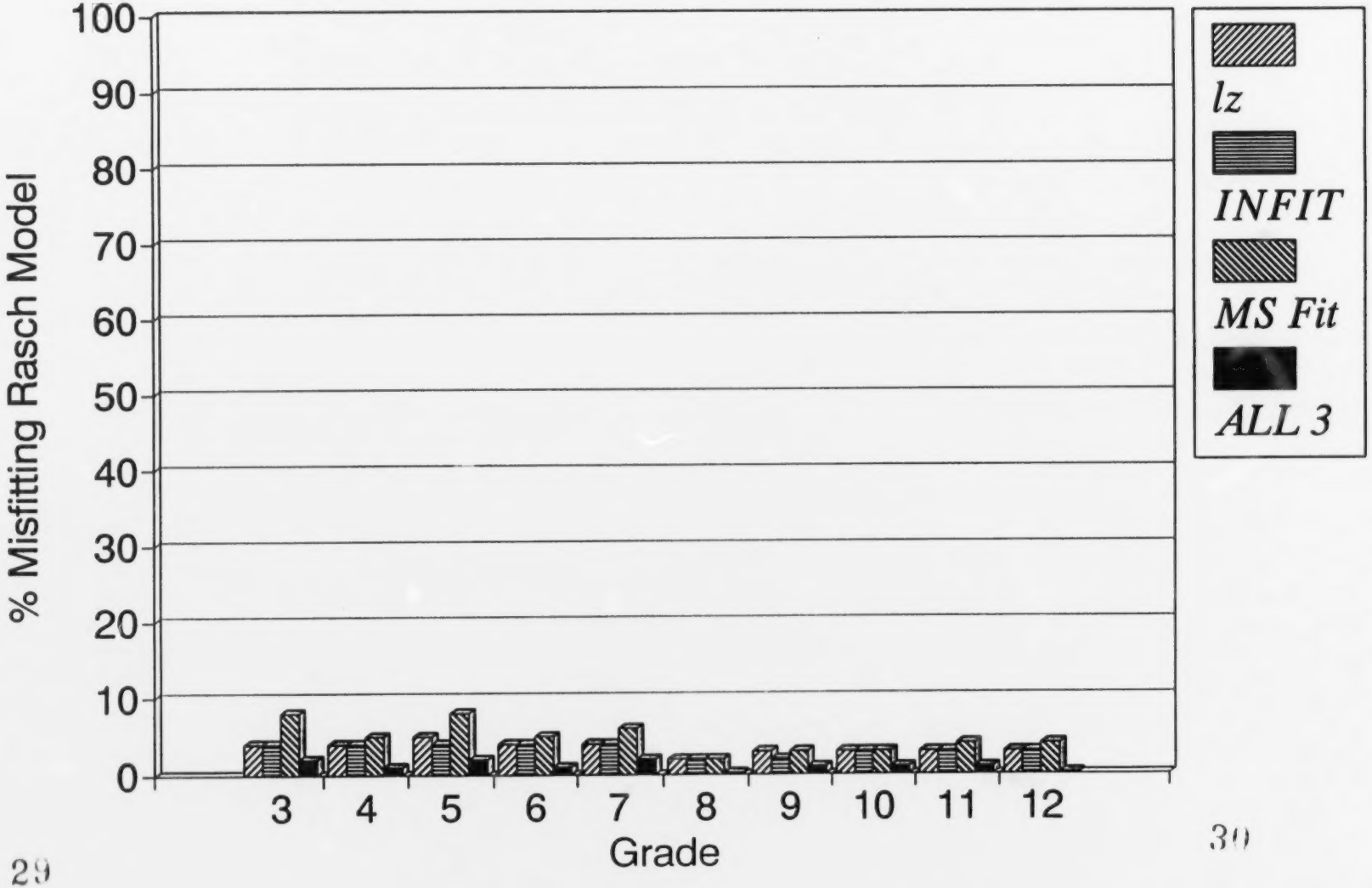


Figure 4
Percentage of Misfitting Students
Reading Comprehension Subtest

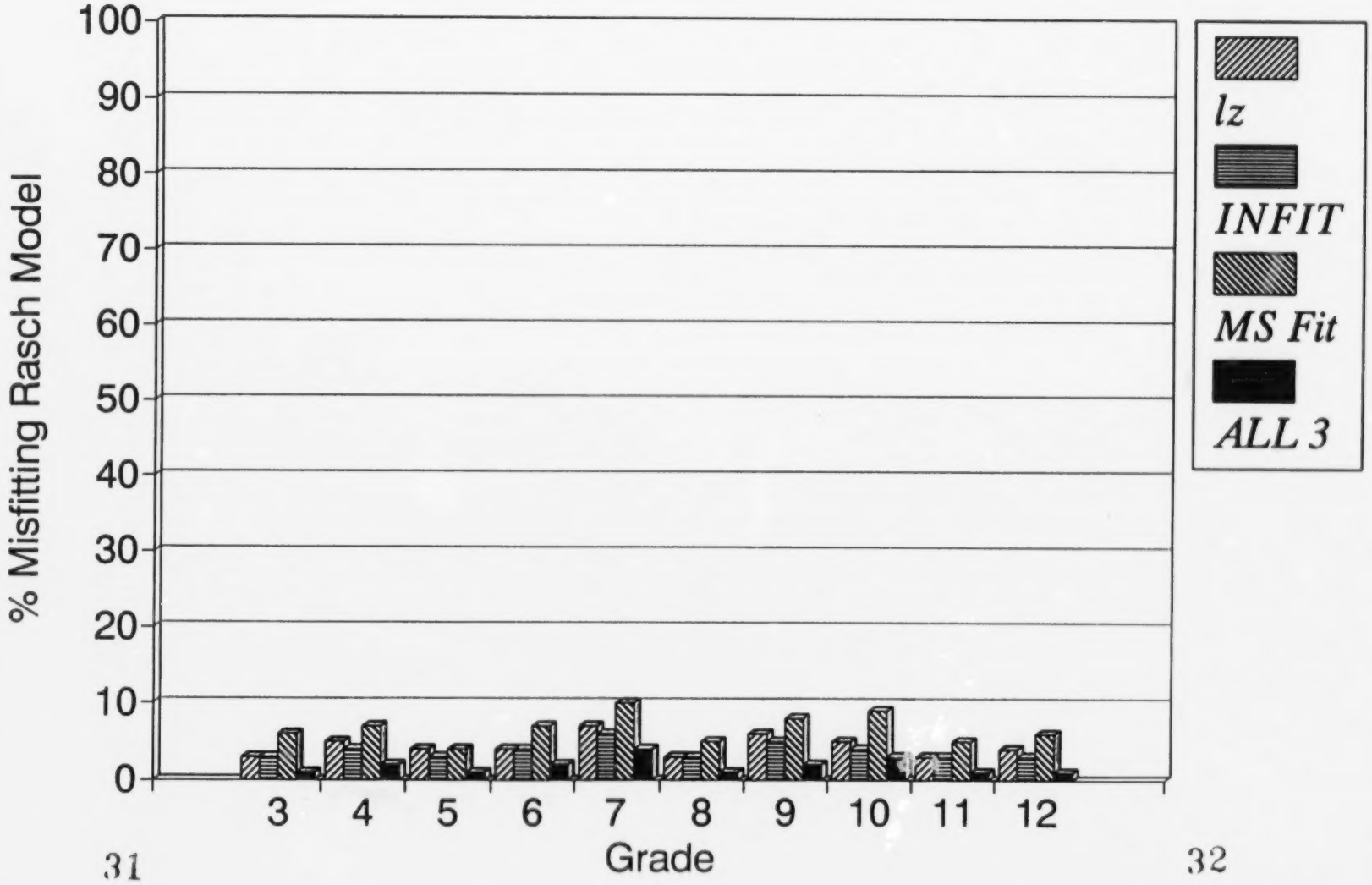
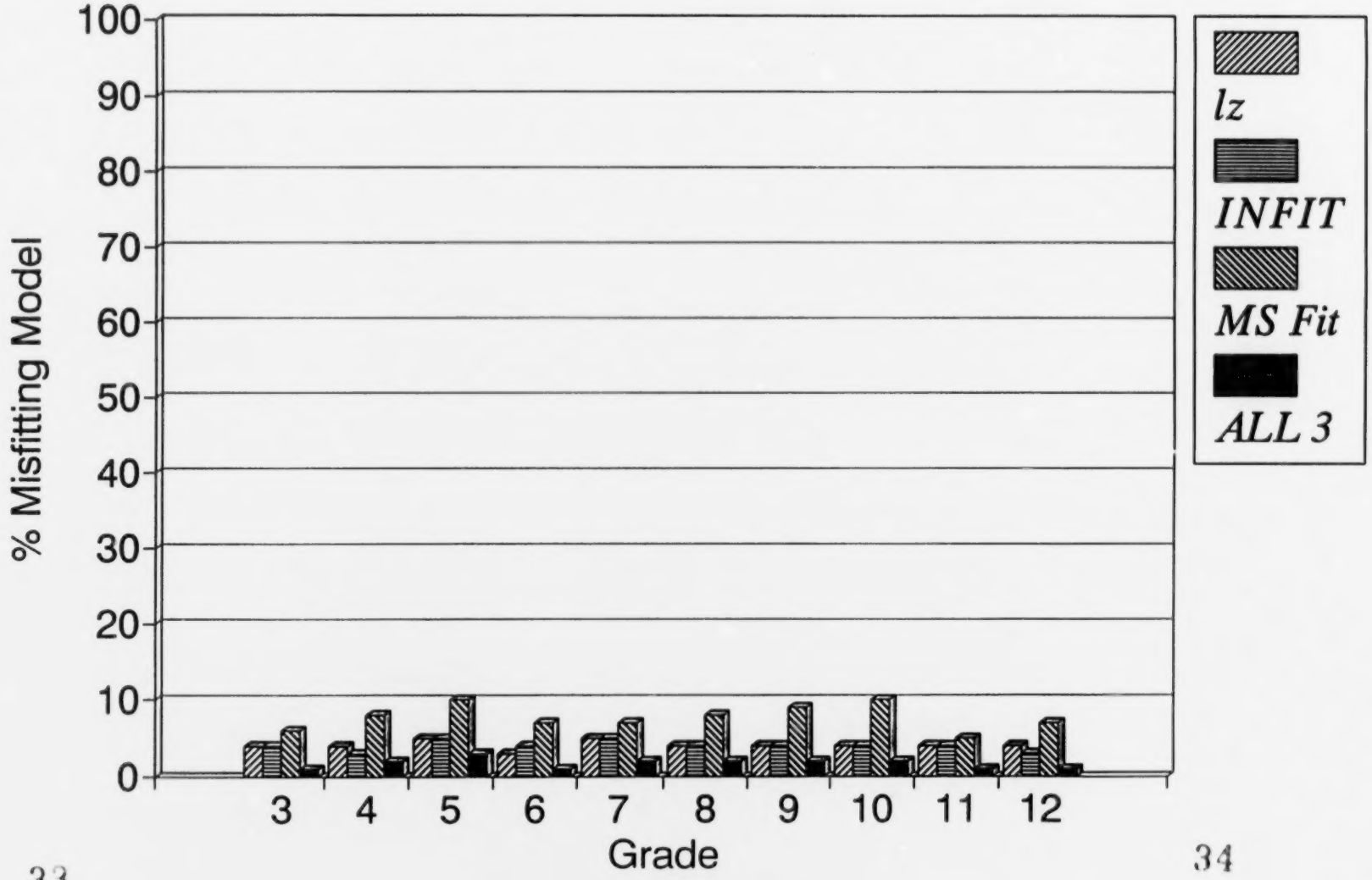


Figure 5
Percentage of Misfitting Students
Spelling Subtest



BEST COPY AVAILABLE